Fred Dickinson fredsterdickinson@gmail.com

Variance and Standard Deviation

Try to understand the fundamental ideas behind variance. If you can, think of your own small examples of data to support the concepts being introduced here.

Statistics are a common and (sometimes)[1] effective way to describe data. For example, the *mean* is the arithmetic average over your data: add it up, and divide by how many data points you have. Consider the following data set x that represents the heights of players on a small basketball team

$$x: \quad 180 \quad 179 \quad 177 \quad 183 \quad 180 \quad 181 \qquad \text{(cm)}.$$

You should verify for yourself that the mean of this data, denoted $\overline{x}$, is 180.

Now consider a slightly different data set; this time the heights of a small selection of commuters on a train. This data set is labelled y and is

$$y: \quad 174 \quad 162 \quad 191 \quad 185 \quad 178 \quad 190 \qquad \text{(cm)}.$$

What is the mean, $\overline{y}$? You should again check for yourself that $\overline{y} = 180$. So, if we are just using the mean to describe these data sets, they are the same...

... but of course they are not. In fact, they are quite different to each other. The first set x is very concentrated around 180cm: the points never go more than 3cm away from it. On the other hand, y has data that is very far from 180 - take, for example, the entry of 162cm.

This is precisely how we describe **variance.** It measures how far, on average, your data points are from the mean. The data set x has a small variance since most of its data is near 180cm, whereas the data set y has a larger variance because its data is more *spread out.*

Let's think about how we can formalise this idea. One way to measure how far your data is from the mean is as follows.

(i) For each data point, determine it's difference from the mean;
(ii) Add up all of these differences, and divide by how many data points you have.

This is the 'average distance from the mean'. Unfortunately this method is inherently flawed. Try it for yourself on both x and y. If you do, you'll find that the total sum of 'distance from the mean' is zero. For example, with the basketball player data,

$$\begin{array}{rcccccccl} x: & 180 & 179 & 177 & 183 & 180 & 181 & & \text{(cm)}, \\ x-\overline{x}: & 0 & -1 & -3 & 3 & 0 & 1 & & \text{(cm)}, \end{array}$$

and $0-1-3+3+0+1=0$. This is not a surprise - the mean is **defined to be** the 'middle point' of all your data, so of course the average distance is zero! The data that lies above the mean (positive) 'cancels out' the data below the mean (negative).

[1] Although, abuse of statistics is widely spread and part of your maths education is to learn how critically analyse the information you're given! There is a very interesting website linked here that I recommend giving a quick scroll through.

So we need another way to measure the distance from the mean. The two most obvious choices are to

(a) take the *absolute difference* from the mean, $|x - \overline{x}|$. For example, we would say that 177 is 3 away from the mean;

(b) or take the *squared difference* from the mean, $(x - \overline{x})^2$. In this case we say that 177 is $3^2 = 9$ away from the mean.

As it turns out we choose option (b). This is because using the square has much nicer mathematical properties[2] than the absolute value.

Consider again the basketball player data set x. Now we will use the squared difference from the mean

$$\begin{array}{rcccccccc} x: & 180 & 179 & 177 & 183 & 180 & 181 & & \text{(cm)}, \\ (x-\overline{x})^2: & 0 & 1 & 9 & 9 & 0 & 1 & & \text{(cm)}, \end{array}$$

so if we take the average over these differences, we will find the variance of x to be

$$(0 + 1 + 9 + 9 + 0 + 1)/6 = 3.\dot{3}.$$

The variance is typically labelled $\text{Var}(x)$, or sometimes $\mathbb{V}(x)$. We will use Var, and in this instance $\text{Var}(x) = 3.\dot{3}$ or $\frac{10}{3}$.

So we have a way to describe how far the data is from the mean - that's great, right? There's still just one issue. We haven't made a mistake, but $\text{Var}(x) = 3.\dot{3}$ despite no individual point in x being more than 3 away from the mean; the furthest data point from 180 is the entry of 183.

We have *lost our scale* by taking the squared distances. The variance is *not measured in the same units* as the original data. To account for this, we introduce the 'standard deviation', aptly named because it standardises the variance (otherwise known as the 'deviation from the mean'). It does so by taking the square root, i.e.

$$\text{standard deviation} = \sqrt{\text{variance}}.$$

This measurement is on the same scale as the data set. For example, the standard deviation of x is $\sqrt{3.\dot{3}} \approx 1.83$. Convince yourself that this is a better 'average' for the distance from the mean.

Some final remarks on notation. The standard deviation is often labelled with σ, the Greek (lowercase) letter 'sigma'. It follows that the variance is denoted σ^2.

Summation notation. Entirely coincidentally, the Greek (capital) letter 'sigma' Σ is used to describe the sum of a collection of items. It does so with a subscript (below) and a superscript (above). The subscript tells you where to start the sum, and the superscript tells you were to finish. For example, suppose

$$x: \quad x_1 = 180, \quad x_2 = 179, \quad x_3 = 177, \quad x_4 = 183, \quad x_5 = 180, \quad x_6 = 181.$$

[2] For example, the square e.g x^2 is *differentiable* everywhere - whereas the absolute value $|x|$ is not differentiable everywhere (in particular, it is not differentiable at $x = 0$). You should see this later on in your A-level.

Then the sum of all these terms can be represented

$$\begin{aligned}\sum_{i=1}^{6} x_i &= x_1 + x_2 + x_3 + x_4 + x_5 + x_6 \\ &= 180 + 179 + 177 + 183 + 180 + 181 = 1080\end{aligned}$$

which we could then average to take the mean, i.e

$$\overline{x} = \frac{1}{6}\sum_{i=1}^{6} x_i = \frac{1}{6} \times 1080 = 180.$$

For a general data set with N points, we write

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i.$$

We can also express the variance in this way. Remember that the variance is just the average of the *squared difference* between each data point and the mean $\overline{x}$. So,

$$\mathrm{Var}(x) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2.$$

There are a few different ways to express this sum, but we will not cover these here.

Exercise 1.1. *Explain, in your own words, what the standard deviation of a data set is. How is it different to the variance?*